DOCUMENT RESUME

ED 170 351 TM 008 766

AUTHOR Spencer, Mary L.; And Others
TITLE Measures of Non-Academic Functional Literacy in Children. An Evaluation of Available Instruments.
INSTITUTION Pacific Training and Technical Assistance Corp., Berkeley, Calif.
SPONS AGENCY System Development Corp., Santa Monica, Calif.
PUB DATE 13 Oct 75
NOTE 96p.; For related document, see TM 008 749

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Basic Skills; Compensatory Education Programs; *Evaluation Criteria; *Functional Illiteracy; *Functional Reading; Intermediate Grades; Junior High Schools; Literacy; Program Effectiveness; *Reading Tests; *Test Reviews; *Test Selection
IDENTIFIERS Adult Performance Level; Basic Skills Reading Mastery Test; Elementary Secondary Education Act Title I; Fundamental Achievement Series; National Assessment of Educational Progress; New York State Basic Competency Test in Reading; Reading Everyday Activities in Life; Test of Adult Functional Competency

ABSTRACT

As part of the development of a functional literacy test for fourth through eighth grade children in Title I compensatory education programs, this report enumerates a set of criteria for selecting appropriate tests. The criteria are grouped into six categories: (1) test background; (2) psychometric quality; (3) examinee appropriateness; (4) normative standards; (5) administrative usability; and (6) interpretation. The six tests reviewed as potential instruments are the Adult Performance Level Test, Basic Reading Skills Mastery Test, Fundamental Achievement Series, National Assessment of Educational Progress, New York State Basic Competency Test, and Reading/Everyday Activities in Life. None of these tests meets all the criteria. Alternative solutions proposed include developing a new test, or constructing a test using parts of existing instruments. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED170351

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

MEASURES OF NON-ACADEMIC FUNCTIONAL LITERACY IN CHILDREN

AN EVALUATION OF AVAILABLE INSTRUMENTS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Mary Spencer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

By

Mary L. Spencer

Nicolas Fedan and Bobby R. Offutt

Submitted to: System Development Corporation
2500 Colorado Avenue
Santa Monica, California 90406

Submitted by: Pacific Consultants
3099 Telegraph Avenue
Berkeley, California 94705

October 13, 1975

TM008 766

ACKNOWLEDGEMENTS

Several test authors, publishers, and agency sources of tests were extremely helpful to this review. Their efforts to relay test materials and information in a highly expeditious manner are appreciated, particularly in cases where materials pending copyright were entrusted to the reviewers. Mention is due to the following persons:.

Dr. Norvell Northcutt
Adult Performance Level Test

Dr. Kenneth Majer
Basic Skills Reading Mastery Tests

Dr. Marilyn Lichtman
Reading/Everyday Activities in Life

Jane Algozine
New York State Basic Competency Test

Dr. Harold Wilson
National Assessment of Educational Progress

Psychological Corporation
Fundamental Achievement Series

MEASURES OF NON-ACADEMIC FUNCTIONAL LITERACY IN CHILDREN

AN EVALUATION OF AVAILABLE INSTRUMENTS FOR INCLUSION IN THE STUDY OF SUSTAINING EFFECTS OF ESEA TITLE I COMPENSATORY EDUCATION

TABLE OF CONTENTS

	Page
INTRODUCTION	1
Requirements for a Measure of Functional Literacy in Children	2
TEST REVIEW CRITERIA	5
Outline	7
Test Background	8
Psychometric Quality	9
Appropriateness	19
Normative Standards	23
Administration	24
Interpretation	25
TEST EVALUATION	27
Adult Performance Level Test	28
Basic Reading Skills Mastery Test	37
Fundamental Achievement Series	45
National Assessment of Educational Progress	53
New York State Basic Competency Tests	66
Reading/Everyday Activities in Life	79
SUMMARY AND IMPLICATIONS	85
REFERENCES	91

MEASURES OF NON-ACADEMIC FUNCTIONAL LITERACY IN CHILDREN

AN EVALUATION OF AVAILABLE INSTRUMENTS FOR INCLUSION IN THE STUDY OF SUSTAINING EFFECTS OF ESEA TITLE I COMPENSATORY EDUCATION

INTRODUCTION

The Study of the Sustaining Effects of Title I Compensatory Education on Basic Skills, to be conducted by System Development Corporation, will describe and evaluate the Title I services for economically and educationally disadvantaged children. The assessment of student performance has been identified as one of the major activities of the study. Toward this end, a norm-referenced evaluation model will be implemented via the administration of a standardized achievement test to 160,000 children in the Fall and Spring of three consecutive school years, beginning in the Fall of 1976. Due to increasing concern regarding the use of standardized achievement tests with disadvantaged and minority students, the United States Office of Education has directed that measures of more *life-like, non-academic,* or *functional* instances of literacy in children be evaluated for selection as an adjunct index of reading and mathematics ability.

In service of the need to supplement indices of reading and mathematics ability derived from standardized achieve-

ment tests, Pacific Consultants has conducted an extensive search and review of relevant literature from educational and psychological journals, research and information clearinghouses, government projects, and individual investigators. This process has resulted in the production of a definition of functional literacy in schoolchildren, the development of criteria for the evaluation and selection of a measure of functional literacy in schoolchildren, and the identification of a set of candidate instruments. While the definition and criteria were delineated in a previous report (August 29, 1975), the purpose of the present report is to set forth a refined set of criteria for test selection, and to review and evaluate the identified measures. Subsequently presented are a summary of findings, and a discussion of their implications for the assessment of functional literacy in the Title I Study of Sustaining Effects.

Requirements for a Measure of Functional Literacy in the ESEA Title I Study of Sustaining Effects

A general description of the characteristics desired in the test of functional literacy was stated in SDC's Statement of Work for the Title I evaluation and amplified in later discussions between SDC, Pacific Consultants, the United States Office of Education, and the Functional Literacy Panel. These characteristics can be briefly represented as follows:

First, the instrument must clearly measure the operational definition of functional literacy that was developed for the Study of Sustaining Effects. Accordingly, functional literacy is viewed as the reading and computational skills needed by children as they deal with the contemporary non-school related world. It must be an independent test in the sense that it was specifically designed to measure functional literacy rather than being the reading or computational portion of an achievement test battery.

Second, the level, range, and content of the test must be appropriate for elementary school children in grades 4 to 8, including children from disadvantaged backgrounds.

Third, costs of the test should be in the normal range of costs for comparable tests.

Fourth, the test must be capable of group administration by non-expert school personnel employing uniform procedures across the country.

Fifth, the test should be amenable to machine scoring.

Sixth, if a norm referenced test is used, the norms pertaining to the population of the study should be available. If a criterion-referenced test is used, the criteria on which the test is developed should relate in a valid way to compensatory education objectives.

Seventh, evidence of reliability should be available.

Eighth, the test must validly measure the behavior addressed by the operational definition of functional literacy.

Ninth, the test must reflect concerns for the propriety of content for a pupil population which is highly diverse in basic skills and ethnic background.

Tenth, the test should effectively meet the purposes of the Title I Study of Sustaining Effects.

TEST REVIEW CRITERIA

Characteristics of the test, the nature of the examinees, and the purpose of testing are important factors in selecting a test of functional literacy for use in the ESEA Title I Study of Sustaining Effects. The criteria for test selection presented here are based very largely on the general guidelines provided by the American Psychological Association's *Standards for Educational and Psychological Tests (1974)*, and the criteria employed in the test evaluations at the Center for the Study of Evaluation (CSE), as presented in the document *CSE Elementary School Test Evaluations (1970)*, authored by Ralph Hoepfner and others. Additional criteria were suggested by Pacific Consultants' previous review of reading and literacy tests for the Right-to-Read Evaluation, (1974), and the recent examination of tests of adult functional literacy performed at the Northwest Regional Education Laboratory under the direction of Dean Nafziger (1975).

The criteria suggested by the sources indicated above provided a reasonably complete compilation of factors relevant to test selection, but were not concerned specifically with the measurement of functional literacy in grades 4 to 8 for the purpose of program evaluation. A number of general recommendations were not suitable in meeting the special requirements of the Title I Evaluation, and were

therefore modified as necessary. The proposed criteria are organized according to six general areas: 1) test background; 2) psychometric quality; 3) examinee appropriateness; 4) normative standards; 5) administrative usability; and 6) interpretation. These are in approximate correspondence with the areas identified in the CSE test evaluation system.

OUTLINE AND WEIGHTING OF TEST REVIEW CRITERIA

Test Background

Criteria in this section consider whether the purpose for which the test was constructed is clearly stated, and whether the manner of test construction is compatible with its purpose.

Psychometric Quality

1. Validity

   Content validity includes criteria encompassing a clear definition of what the test measures; criteria on the behavioral, linguistic, socio-economic function, program referencing, and task-reference of the test items. Empirical validity includes criteria on the empirical evidence relating test scores to other variables or real-life outcomes. Construct validity includes criteria on the theoretical basis of the functional literacy concept.

2. Reliability

   Criteria are included for four measures of reliability: comparability, or alternate-form equivalence; stability, or test-retest correlations; internal consistency, or item intercorrelations; and standard error of measurement, or how much a score is likely to vary.

3. Test-Item Structure

   Included are criteria on the relevance of item construction procedures, item selection procedures, and item difficulty.

Appropriateness

1. Instruction

   Criteria are included for insuring that instructions are clear, that they include an unambiguous explanation of the purpose of the test, that they are comprehensive, that sample item(s) are included, and that they be presented orally.

2. Items

Criteria are included in this section which insure that items have propriety, and be motivating to examinees.

3. Format and Procedure

Criteria here address physical quality, lay out, timing, response mode, and complexity of the test.

Normative Standards

Criteria in this section address availability of norms, quality and representativeness of the normative sample, reporting categories, and desirable types of item statistics.

Administration

This area addresses desireable characteristics of the test in terms of examiner expertise, optimal length of testing time, scoring, test setting, materials, and cost.

Interpretation

Criteria of interpretation address the quality and organization of manuals, clarity of score interpretation, and the implication of the test score.

Test Background

1. The purpose of the test should be explicitly stated. For instance, "...to measure the examinee's performance on tasks or activities held to be significant to a student's life outside of school." Examples should be mentioned (e.g.,"...to read and follow the directions on a medicine bottle").
2. The purpose statement should be clear to those individuals likely to administer the test.
3. The construction of the test should follow closely the purpose for which the test was built. Thus, a diagnostic test (such as the one envisioned for the present study) should state how the test's purpose translates or agrees with the scope of tasks and operations to be covered. Such scope should be limited, well defined, and detailed.

   Diagnostic tests should be designed to yield scores on the separate components of interest. Since the test is not envisioned to be a selection or certification instrument, the range of item difficulties and the power with which the test discriminates among examinees is less important. (A

possible variation from this criteria would arise if the test is envisioned to discriminate among age groups. In this case, the test would also have a classificatory purpose which requires different item properties).

4. The test should indicate whether differences among minority groups were considered during test construction. If such consideration was exercised, items would have been sampled which depict the actual behaviors of students in these groups in extracurricular life activities.

Psychometric Quality

Criteria addressed in this section pertain to the validity, reliability, comparability of the test scores, and the quality of normative standards.

1. Validity -- The criteria in this area concern the nature of what is measured by the test. It is most important that the test be clearly and unambiguously a measure of *functional literacy*, if its role in the Title I Evaluation is to be served. Factors contributing to the credibility of a test as measuring functional literacy are considered in terms of content, empirical, and construct validity.

a. Content Validity -- It is highly desireable that the test be representative of a defineable population of items and performances with specific reference to the domain of functional literacy. The bases of definition and the procedures of test construction contribute to content validity in terms of the criteria outlined below.

1) Definition -- The test should be specifically designed as a test of functional literacy. Disagreement on the validity of content will surely arise if the test was originally designed for some other purpose, and if no explicit basis exists for judging the relevance of items.

2) Material Domain -- The stimulus materials should be representative of those commonly encountered in real-life reading and computational tasks. Confidence in the representativeness of materials would be increased if a population of such materials were defined, the composition of the population was described in terms of types or characteristics of materials, and formed part of the definition of functional literacy used as the basis of test development.

3) Behavior Domain -- The performance required in the items should be as close an approximation as possible of the tasks and skills commonly required in real-life reading and computational performances of children between the ages of 9 to 14. Explicit classification and/or description of a domain of functional literacy behaviors is desireable as part of the definition used as a basis for test development.

4) Symbolic Domain -- The language and other symbolic representations which form the communicative component of the materials should be representative of the symbolic content commonly encountered in real-life reading and computational tasks. Specification of the symbolic content in linguistic and mathematical terms can further strengthen and clarify the definition of functional literacy beyond the material and behavior specifications usually considered. Such specifications could be particularly helpful in defining levels or ranges of competence in relation to the domains of materials and tasks.

5) Socioeconomic Domain -- The materials and tasks should be representative of the socioeconomic functions commonly encountered in real-life reading and computational tasks.

A classification or description of socioeconomic functions, and the benefits or values of performance, should be part of the definition of functional literacy that is used as a basis of test development. Such a classification or appraisal system would help insure that functionally significant, rather than trivial performances are represented.

6) Program Objectives -- The materials and tasks of the functional literacy test should *not* be referenced to specific program objectives. Program-referencing would amount to prejudging the result of the evaluation in relation to functional literacy, in that it would inevitably bias the evaluation in favor of program goals and those programs which emphasized the defined objectives. The test is intended to provide an objective criterion by means of which the effectiveness of various programs can be judged in the area of functional literacy.

7) Criterial Objectives -- The definition of functional literacy should be supplemented and operationalized by the specification of a set of criterial tasks referenced directly to the characteristics of materials, behavior, symbolic content, and functions employed in

the functional literacy definition. Such objectives would provide an important link between definition and items. Such objectives might be used in item construction and selection, or as a basis for empirical validation of items.

b. Empirical Validity -- It is desireable that the test have been used in previous studies, thus providing empirical evidence relating the test scores meaningfully to other variables. Areas of concern in relation to empirical validity are outlined below.

1) Concurrent Relations - It is advantageous but not essential that the test has been correlated in previous studies with a wide variety of other measures taken at the same time. The number and quality of studies, the number of variables, and the diversity of variables all contribute to the evidence bearing on the meaning of a given literacy score.

2) Predictive Relations - It is advantageous but not essential that the test be correlated with measures taken at some later time. The number and quality of studies as well as the number and diversity of variables contribute to the evidence bearing on the

question of what consequences flow from having attained a given literacy score.

3) Causality - It is advantageous, though not essential, that studies have been performed which relate the functional literacy test to important psychological, educational, or socioeconomic independent variables. Such evidence should be of assistance in the analysis and interpretation of the findings in the Title I Evaluation.

4) Nature of Relations - Empirical relationships found in the available literature should be reasonably interpretable in terms of prevailing educational, psychological, and socioeconomic theory. The measure of functional literacy should relate sensibly to variables which can be considered to reflect components of functional literacy, and to variables which are thought to be independent of functional literacy. Factor analytic studies, if any are available, should indicate that the measure of functional literacy is factorially complex. The nature of one particular relationship is especially important. The subtests should not correlate too highly with standardized tests of reading ability or pure computational skills. Very high correlations

of this sort would indicate that the test did not adequately represent the separate skills required in a functional literacy measure.

5) Sensitivity - It is advantageous that the magnitude of effects observed was substantial when the test was used as a dependent variable in experiments or evaluations. That is, the test should be sensitive to the effects of appropriate independent variables, so that there is some assurance that appropriate effects will be revealed in the Title I Evaluation as well.

c. Construct Validity -- Criteria in this area have to do with the theoretical basis of the functional literacy concept. They are of lesser importance in judging validity than are content and empirical criteria, given the practical concerns of the Title I Evaluation. But, they are valuable characteristics nonetheless.

1) Process Constructs - The conceptualization, development, and empirical validation of the test should be grounded on relevant psychological, linguistic, educational theory in the area of reading and computation. Particularly important in this respect is the

availability of a task-skills analysis which would define the components of functional literacy, indicate hierarchical relations among components, and the relationship of performance to basic cognitive information processing operations. Such a theoretical foundation is useful in generating hypotheses and interpreting results.

2) Acquisition Constructs - The conceptualization, development, and empirical validation of the test should be grounded in relevant psychological, linguistic, and educational theory in the areas of instruction and cognitive and language development. Such formulations would provide a basis for tying changes in functional literacy to specific educational practices, and related developmental changes.

3) Socioeconomic Constructs - The conceptualization, development and validation of the test should be grounded on relevant social and economic theory to provide a basis for hypothesis and interpretations of findings concerning relevant socioeconomic variables, and the function and benefits of literacy.

2. Reliability -- The question here is how well does the test measure what it does measure?

a. Comparability - If alternative forms are available, they should be based on parallel items with comparable item statistics. The forms should correlate .80 or above at every grade level in the 4 to 8 grade range. Although seldom provided in the early stages of test development, this is the preferred measure of reliability. In practice, two forms would be considered comparable (equivalent) if, 1) they include the same number and kind of items; 2) standard deviations in the two forms are not significantly different; and 3) means obtained with the two forms are not significantly different.

b. Stability - Test-retest correlations should be .80 or above over brief time intervals; i.e., one month or less. Reliability coefficients could be lower over longer intervals, particularly when instructional experiences have intervened, having a substantial effect on the level of functional literacy performance. However, in the case where no shift in level of performance has occurred, the reliability should remain above .70 for intervals up to one year.

c. Internal Consistency - High internal consistency is not a *necessary* criterion for the functional literacy test, since a test which is highly

homogeneous is not likely to represent the full diversity of tasks which should be sampled in a functional literacy test. In particular, items involving reading should only be moderately related to computational items. The correlation between reading and computational subtests, if present in the test, should correlate below .70, and preferably below .50. Where alternate forms are available, then evidence of internal consistency is highly desirable.

d. Standard Error Of Measurement - A statistic which allows an interpretation of the reliability of each score is desireable. If the test discriminates at various age levels, the standard error of measurement for each level would show how well this differentiation is accomplished.

3. Test - Item Structure

a. Item Construction - Procedures used in item sampling should be clearly defined and replicable. It is necessary that test information indicate the relevance and representativeness of the item pool in relation to the aspects specified in the definition of functional

literacy, whether material, behavioral, symbolic, or socioeconomic criteria are included. Procedures which are entirely algorithmic would be most advantageous but are not within the usual state of the art at present. Other procedures are acceptable if the resulting items show close correspondence to the classification systems employed in defining functional literacy.

b. Item Selection - Procedures used in selecting items from a pool for inclusion in the final test should be based on observations of actual behavior, and yield evidence that the items load evenly on the various categories defining functional literacy.

c. Item Difficulty - Items should include a wide range of difficulties, including some items relatively easy for 4th grade children, and some items relatively difficult for 8th graders. Additionally, in view of the diagnostic use of the test, it should include a sufficient number of "easy" items so as to yield a useful analysis of examinees' strengths and weaknesses.

Appropriateness

The third set of criteria concern the appropriateness of the test in relation to characteristics of the intended

sample of examinees. The criteria focus on the three areas of instructions, items and format, and procedure. The present criteria insure that irrelevant sources of difficulty are eliminated from the test.

1. Instructions

   a. Clarity - The instructions should be appropriate in orientation and tone, inoffensive in content, and comprehensible, with vocabulary and syntax suitable for children in the 4 to 8 grade range.

   b. Purpose - The instructions should provide an honest explanation of its purpose and intended use.

   c. Comprehensiveness - The instructions should precisely and completely describe all requirements of the tasks presented in the items so that the examinee has all the information needed to adopt an effective performance strategy. Appropriate instructions should be included on the relation between guessing and test scores.

   d. Sample Items - The instructions should include sample items accurately illustrating task requirements and the level of difficulty of the tasks.

reading and computational tasks should present realistic facsimiles of the actual materials,

e. Mode - The instructions should be presented in an oral mode. A standardized script should be available which is suitable for fluid oral reading by non-expert examiners.

2. Items

a. Motivation - The items should be relevant, up-to-date, and interesting for children in the 4 to 8 grade range, so as to arouse intrinsic motivation in task performance without extensive exhortations being required to induce cooperation and effort.

b. Propriety - The content of the items should not involve any invasion of privacy, or any sexist, racist, or otherwise offensive aspects of content.

3. Format and Procedure

a. Physical Quality - The paper should be of good quality, the print bold and readable, and the illustrations clear and up-to-date. Reproduction of materials involved in common

preferably including full-color reproductions.

b. Layout - The test should be effectively arranged and cued to facilitate recognition of items as units, the perception of the relation of item stems to answers and examinee response, and the progression of successive items and pages.

c. Timing - The test should be time limited but permit most examinees to attempt most items within the time allowed. Sectioning of the test, with timing instructions for each section may help to maintain appropriate pacing in the brief time alloted for this test. Items at all difficulty levels should be represented in each section.

d. Response Mode - The response should be marked in a fashion permitting machine scoring.

e. Complexity - Each item should require one simple and direct response, with no multiple steps or complications other than those intrinsic in the task represented by the item. Several items might be used based on

the same stimulus materials, provided that the relationship of each item to the stimulus is clear.

Normative Standards

1. Data Available. Although normative data is not essential in view of the large sample to be tested in the Title I Evaluation, and the emphasis on program comparison in the evaluation, it will still be helpful to have some prior normative data available as a basis for comparison.
2. Normative Sample. It is desireable that normative data be available for the 4 to 8 grade range, and for adults as well.
3. Representative. It is desireable that the sample be representative of racial, ethnic, sex, geographic, and socioeconomic strata, rather than the result of incidental sampling.
4. Reporting. It is desireable that normative data can be reported separately as well as combined over the racial, ethnic, geographical, and socioeconomic strata represented in the sample.
5. Item Statistics. It is useful if item statistics are reported both for the whole sample and for the separate strata. Item difficulties are

the most important statistic, but if selection or classification uses are envisioned, then item discrimination indices and intercorrelations are useful as well.

Administration

1. Personnel. Non-expert school personnel should be capable of administering the test with very little training. The services of a specialist or a testing expert, or extensive training should not be required.
2. Scheduling. The test should require no more than 30 minutes of testing time (preferably 20 minutes) on one occasion of testing. Tests taking longer than 30 minutes should be easily modifiable for a shorter length, with no more than normally expected loss of reliability.
3. Setting. The test should be capable of administration in usual classroom settings, to group sizes in the normal range for intact classroom groups, and without the necessity of special equipment.
4. Scoring Method. The test should be scored in an objective manner by machine. Machine scoring should be highly fail-safe and reliable,

without complex error checking routines to proof the results.

5. Materials. - The test materials should be entirely of the paper-and-pencil test variety, with no special manipulanda, slides, or other unusual components.
6. Cost. - Costs should be in the normal range of paper-and-pencil tests having good quality paper and printing, including color reproduction.

Interpretation

1. Manuals. - A high quality test manual should be available, one which meets the appropriate APA standards (1974) for test manuals. A supplemental brochure describing the test and how to interpret its scores should also be available for relatively unsophisticated consumers of the results.
2. Meaning. The test scores should be highly meaningful and understandable in terms of specific performance by a nontechnical audience including the general public. It would be most meaningful if a hierarchy of performance levels could be devised, in which a person placed at one level could be

described as being capable of a specific list of tasks, and all tasks listed at lower levels. However, this may be an unrealistic goal.

3. Scales. The primary test scores should be directly understandable in absolute terms without the use of complex conversions or scaling. Forms of scaling or conversion to standardized scores may be used as a supplement to the primary scores or for use by audiences with a higher level of technical background.

4. Implications. It is desireable that the implications of given test scores for educational practice or public policy be clear and relatively direct. However, what is actually required to meet this criterion is not entirely certain.

TEST EVALUATION

Six instruments were selected for review according to the test review criteria: The Adult Performance Level Test, the Basic Reading Skills Mastery Test, the Fundamental Achievement Series, the National Assessment of Educational Progress, the New York State Basic Competency Test, and the Reading/Everyday Activities in Life. Three of these instruments were reviewed previously by Northwest Regional Educational Laboratory under the direction of Dean Nafziger (1975). However, the focus and criteria of that review were determined by a different set of purposes and population characteristics than are operational in the Title I Study of Sustaining Effects. These and the remaining three tests reviewed in this report were selected on the basis of the reviewers' judgement that they possessed some property or set of properties that placed them within the range of promising instruments for the Title I study's purposes. It should be clear from the outset that the judgements made of these tests relate only to the potential usability of the instruments in the Title I Study of Sustaining Effects, and the test evaluations should not in any way be construed as either indictments or recommendations of the instruments for adoption in other contexts.

Adult Performance Level Test of Adult Functional Competency

Adult Performance Level Project
Dr. Norvell Northcutt, Project Director
The University of Texas at Austin
Division of Extension
Austin, Texas 78712

Description

The Adult Performance Level (APL) Project of the University of Texas Division of Extension developed the APL test as part of their mission to, ". . . specify the competencies which are functional to economic and educational success in today's society and to develop devices for assessing those competencies of the adult population of the United States." The instrument is currently in an experimental form and is not considered by its author to be ready for utilization. It is a short form test of 42 questions that are related to a variety of adult life experiences. For example, items include a 1040 Individual Income Tax Return, a bank deposit slip, itemizing of grocery bills,and tax deductions. They require performance in the areas of communication, computation, problem-solving, and interpersonal relations.

The APL field-test data were used to define three functional categories: 1) adults who function with difficulty; 2) functional adults; and 3) proficient adults. Each of the three APL levels is based on three criteria: 1) predicted income; 2) education; and 3) job status. The people in the first APL category are considered to be functionally incom-

petent or to function with difficulty. Those in the second category are competent or functional on a minimal level, and those in the third category are proficient in that they demonstrate competence or that it is associated with a higher level of income and education.

Test Background

1. Purpose

The purpose of the APL test is to measure the competencies of adult Americans which are functional to economic and educational success in today's society.

2. Clarity of Purpose to Examiners

The APL does not have an examiner's manual in which purpose would be explained.

3. Compatibility of Purpose and Test Construction

The APL theory of functional competency was arrived at by focusing on the basic requirements for adult living. A review of the behavioral and social research literature was made in an effort to find a way of categorizing the needs of the undereducated and underemployed adult. The APL project surveyed the State and Federal Agencies in order to select the characteristics that identified the successful from the unsuccessful adult. Additionally, conferences on adult needs were conducted in different regions of the country. Through this process, the APL project developed a multi-faceted model of competency. The elements

of this model included: 1) functional competency as a construct which is only meaningful in a specific societal context; 2) functional competency as a set of skills related to a set of general knowledge areas imposed by society; and 3) functional competency as a dynamic rather than a static process.

The information used to develop the model of functional competency formed the basis of a set of objectives which were then used as a basis for the description of behaviors believed to be important to adult competency. The performance indicators, or items, were written for each competency. This evolutionary process established a close relationship between the purpose and construction of the APL test.

4. Compatibility of Purpose with Item Sampling

The information on item sampling indicates that efforts were made to select tasks which adults actually encounter. There is no evidence that actual observations of adult *competency* behavior were made to verify or generate these tasks. An examination of the tasks shows them to be reasonable experiences for adults, but highly irrelevant to children in the 4th to 8th grade age group.

Psychometric Quality

1. Validity

   a. Content Validity - The APL was defined specifically in terms of the domain of non-academic functional literacy in adults. However, this domain does not correspond well with that of children in the age group to be sampled in the Title I study. The stimulus materials are representative of those commonly encountered by adults in real life reading and computational tasks. There is no evidence that these were sampled directly from the actual universe of adult behaviors. It was not within the APL functional competency study's purpose to even consider the behavior domain of children. The symbolic domain was considerably too complex for the Title I age range.

   The APL test was specifically designed for people who were socio-economically and educationally poor. As desired, the APL was not referenced to a specific set of program objectives. The skills and knowledge areas of the APL competency model overlap to an incomplete extent with the definition of functional literacy adopted for the Title I Study of Sustaining Effects.

b. Empirical Validity - No evidence was presented in the APL materials reviewed which indicated that the APL had been related to measures taken concurrently or at a later point in time. Neither was there evidence to relate the APL to other psychological, educational or socioeconomic independent variables. No information was available on the nature of relations between the APL and other measures, or of the sensitivity of the tasks.

c. Construct Validity - No information was provided.

2. Reliability

When administration procedures are held relatively constant, the relationship of APL performance in two independent samples was highly reliable across various subject characteristics such as income, education, occupational status, urbanicity, ethnicity, sex, age, and other demographic variables. Other indices of reliability were not reported.

3. Test - Item Structure

The item construction of this test, while not traditional, was identified with what could be considered typical adult life experiences and distinct tasks were assigned to each experience. Four primary skills were considered indicative of requirements

placed on adults: 1) reading, writing; 2) computation; 3) problem-solving skills; and 4) interpersonal relation skills. The first two correspond to portions of the definition of functional literacy adopted for the Title I study. The knowledge areas of the APL resemble some of the areas of life activity and the socioeconomic functions included in the operational definition of functional literacy in the Title I study. Item selection was based on expert judgement and revised successively on the basis of field testing. Data on item difficulties were not available. In the reviewers judgement, most APL tasks are much too difficult for children in the Title I age range.

Appropriateness

1. Instructions

In the reviewer's judgement, the vocabulary was somewhat over-sophisticated for the sample of adults taking the test, but consideration was given to making the test less difficult so that every respondent could attempt every task. Children in grades 4 to 8 would probably find that the vocabulary and syntax of the instructions do not correspond with their common life experiences. The task instructions were sufficient for understanding. No sample items were provided. Information on test administration procedures was

unavailable. The instructions are amenable to oral presentation.

2. Items

Most items do not appear sufficiently relevant or interesting to be motivating to 4th to 8th grade children. No offensive content was identified in the APL tasks.

3. Format and Procedures

The physical quality and layout of APL tasks were of superior quality. Many of the stimulus materials were good facsimilies of real-life forms and literacy prototypes. Details on timing were unavailable. Multiple choice and brief examinee-supplied answers were the response modes used. Although single direct responses were usually required in the tasks, some tasks required several examinee-supplied answers.

Normative Standards

Data on the APL were obtained from "representative" samples of American adults. Prior to this, the APL was field tested with 3,500 undereducated and underemployed adults in 30 states. The test has been administered to five independent samples of 1,500 or more persons, for a total of 7,500 adults. Detailed test results were not included in the report made available to the reviewers. The results were used, however, to estimate the percentage of task performances

which could be classified according to the three APL levels of competency for each area of knowledge and in each skill domain. Children were not tested, and no statement was made to indicate that the sample was systematically stratified for various demographic variables.

Administration

The APL tasks are amenable to group administration in a classroom setting by nonexpert personnel. A set of these tasks could be selected in a manner which would produce a 30 minute testing period. No information on the scoring process was available. In the reviewer's judgement, a separate machine scoreable response sheet would be required. The APL materials are entirely of the paper-and-pencil variety. Information on cost was not provided. The test is still in an experimental form.

Interpretation

No test manual was available to the reviewers. It is presumed, though not explicitly described in the material available, that task scores can be linked to the three levels of APL competency. The principal implication of APL task performance pertains to the respondent's ability to perform specific tasks which are generally agreed to have socioeconomic benefits to American adults.

Evaluation

Although the APL measures both reading and computational skills, it is intended for use with an adult population. In the reviewer's judgment, most APL tasks are much too difficult for children in the Title I age range.

From a test-construction point of view, although efforts were made to select tasks actually encountered by adults, there was no evidence that observations of adult behavior were made so as to verify these tasks. The absence of validity data does not allow a judgment on the relationship between the APL and other tests and variables important to the concept of functional literacy.

Modification of this test would depend on the availability of easier items. Since the APL was constructed for use with adults, it is unlikely that enough easy items can be obtained from the item pool, so as to construct an instrument appropriate to the population envisioned in the Title I study. If this process were attempted anyway, a clinical pretest of the new instruments would be necessary before any judgment could be made regarding the tests's appropriatness.

Basic Skills Reading Mastery Test

Maryland State Department of Education
Division of Instruction
Baltimore-Washington International Airport
P.O. Box 8717
Baltimore, Maryland 21240

Description

The Basic Skills Reading Mastery Test consists of three separate forms which purportedly test two of the State's five reading goals with children in three age groups (12 years to adult). Four scored subscales were developed on the basis of behavioral objectives flowing from these goals: 1) following directions; 2) locating references; 3) gaining information; and 4) understanding forms. The test also assesses how students feel about reading.

Test Background

1. Purpose

The purpose of the BSRM test is to asses two of five reading goals adopted by the Maryland State Board of Education. Specifically, these goals were, ". . . to meet the reading demands for functioning in society," and ". . . to select reading as a personal activity."

"Functioning in society" was listed as having five basic goals: 1) following directions; 2) locating references; 3) attaining personal development; 4) gaining information; and 5) understanding forms.

The specific age levels designated as the

National Assessment of ... N.Y. State Basic ... Fundamental

target population were: 12 year-olds, 15 year-olds, and 18 year-olds.

Although the purpose of the test is compatible with the instrument envisioned for the present study, the ages of the target population are only partially overlapping with those to be included in the Title I study.

2. Clarity of Purpose to Examiners

Clarity was good. The manual makes the purpose explicit.

3. Compatibility of Purpose and Test Construction

The five basic goals listed above were translated into specific behavioral objectives by a group of reading and test development specialists. Using these behavioral objectives as a guide, Maryland State department personnel and test developers solicited published materials and printed forms from tax offices, welfare agencies, Chamber of Commerce and other Federal, state and local agencies. These materials were used to generate a bank of "over 500 test items to correspond to the approved series of behavioral objectives. . ." Each of the items was then reviewed by a panel of reading and test-review experts. Student evaluators were also used to assess the items for "clarity, logic, difficulty and readability."

The basic goal of "attaining personal development" was, "... designed to assess the attitudes of students and how they feel about reading." A separate development procedure was employed for this category, with items generated based on the behavioral criteria and then reviewed by teachers and reading specialists. A small-scale field test was performed to refine test length, format, to clarify directions, and to select the best items. Finally, a statewide sample of 2,100 students were selected for the field-test, representing geographical regions and minority groups.

At first, the final test contained basic or easy items, and advanced or difficult items. Later (after the test had been used with 47,000 students), two major changes were incorporated: 1) the distinction between basic and advanced items was dropped in favor of a distinction between items measuring "survival" skills vs. those not so viewed; 2) the test was lengthened to include enough items such that diagnostic information could be obtained for each basic goal.

The present reviewers view the above-mentioned changes as highly compatible with qualities of the instrument sought. However, some problems still

exist as evident from test construction procedures.

4. Compatibility of Purpose with Item Sampling

The manner in which the stimulus materials were obtained for item construction did not include actual observation of the children and young adults for whom the test was intended. Thus, it appears that whereas some items certainly test the survival skills of an adult (i.e., voting directions, applications for driver's license, working permits, W-2 forms, welfare forms, etc.), they do not seem applicable to all ages. In the 15 and 18 year-old forms, some items were present (directions for sewing, an application for U.S. savings bonds, a chart rating household thermometers, etc.) for which there is no evidence that members of the target population have actually been observed reading them.

There is no evidence that test items depict actual extracurricular life activities of minority groups.

Although the psychometric properties of items may be satisfactory (i.e., items actually discriminate among age groups), it is not clear that diagnostic assessments of examinees produced by this test measures survival skills from the frame of reference of the examinee population. It appears

more likely that any such diagnostic information produced, measures survival skills from the frame of reference of the examiners.

The items measuring "attaining personal development" (e.g., "How would you rate yourself as a reader?" "How do you feel about reading as a spare-time activity?") may be highly sensitive to social desirability. No information in this regard was provided by the manual.

Psychometric Quality

1. Validity

   a. Content Validity - The test was referenced specifically to the domain of functional literacy. However, the population of possible stimulus materials was not defined, and thus it is unclear to what degree the five goals for which behavioral objectives were determined fit a theoretical material domain. The test does not include computational tasks. It is not known to what degree the test samples from the behavior, symbolic, and socioeconomic domains.

   b. Empirical Validity - No data was given in the test manual.

   c. Construct Validity - No information was given in test manual.

2. Reliability

Only internal consistency data was available (Kuder-Richardson 20). All coefficients reported are satisfactory. No stability or comparability coefficients were available. No standard error of measurement was available.

3. Test-item Structure

The procedures used in item construction were clearly defined. The item pool was referenced to predetermined behavioral criteria, rather than to various aspects specified in a definition of the possible material domain from which items could be drawn. Item selection was not based on observations of actual behavior. No computational items were included. Item difficulty appears appropriate if referenced to the psychometric properties of the items. But, as mentioned above, it is unclear whether the items sample actual behaviors.

Appropriateness

1. Instructions

They are clear, but do not explicitly state the purpose and intended use of the test. Although they are comprehensive, there are no item examples included. Although some instructions are presented orally, the student is required to read the item instructions. The reading ability required may be more advanced than necessary.

2. Items

To the extent that they sample actual behaviors, they appear relevant and motivating. The propriety of items is maintained throughout.

3. Format and Procedure

The test failed to represent facsimiles of actual materials, thus rendering the physical quality of the test rather poor. Layout, timing, response mode and complexity are acceptable.

Normative Standards

No normative data is available.

Administration

The criteria for this area are deemed to be met by this test. Personnel, required, scheduling, conditions of testing, scoring, and test components are appropriate. Cost of test is not available in the manual.

Interpretation

No separate manual is provided. Beyond referencing a "passing" score (80% of items correct), no mention is made of the meaning of scores or how to interpret them either for diagnostic or achievement purposes. No implications of scores are provided.

Evaluation

Two equally large problems exist with this test from the standpoint of use for the Title I study. First, no computational skills are included; the test was constructed

to be used with a population aged 12 to 18 years old. Thus, it only partially overlaps with the examinees envisioned in the Title I Study.

From a conceptual point of view there are also problems, the most important of which is that the manner in which stimulus materials were obtained for item construction did not include actual observation of the people for whom it is intended. Consequently, some of the items (i.e., questions about an application for a U.S. savings bond) are very unlikely to be appropriate to the population to be tested in the Title I study.

Modification of this test would hinge on the availability of an item pool from which easier items could be obtained; i.e., items which are more appropriate to younger children. In addition, a computational subtest would have to be either constructed or adapted from some other source.

Needless to say, a clinical pretest on the newly-created subtests would be a requirement before any judgment can be made of the test's appropriatness to the Title I Study.

Fundamental Achievement Series

The Psychological Corporation
757 Third Avenue
New York, New York 10017

Description

The fundamental Achievement Series (FAS) was designed as a "culture-relevant" test for the disadvantaged. It consists of a Verbal test and a Numerical test, each requiring a 30 minute testing period. The test is intended for use as an employment, placement, or diagnostic test for adolescents and adults who have had less than the usual exposure to formal education. It yields three scores: Verbal, Numerical, and Verbal + Numerical.

Test Background

1. Purpose. The FAS is oriented toward the measurement of verbal and numerical skills, and is "... intended for use in the employment of adults and adolescents who may not have had the usual exposure to formal education." The test is viewed as a placement and/or diagnostic device. The test cover a range of ability that extends to "... somewhat above the Eighth-grade level."
2. Clarity of Purpose to Examiners. The manual makes the purpose explicit.
3. Compatibility of Purpose and Test Construction. No data on test construction is given in the manual, therefore it is impossible to rate this test on the

relevant criteria. A brief description was given of types of items included, but no mention was made of how these items were obtained. The test has two alternate forms.

Psychometric Quality

1. Validity.

   a. No information is given in the manual on content validity.

   b. Empirical validity. The test was administered to Black employees in a Southern Hospital. Concurrently, these employees were rated by their supervisors on four performance factors. Similar studies were conducted for employees of an Eastern Bank who were in a private-sector anti-poverty training program. In this case, success criteria were teachers' ratings in various subject matter. Six other similar studies were reported in the manual. Correlation coefficients ranged from a high of .62 to a low of -.01. Some of the high coefficients were encouraging in that they actually related to "real-life" criteria. About half the coefficients presented would be considered "useful" in an industrial selection situation. Although the purpose of the present review is to find a diagnostic rather than a classificatory tool, it is to the present measure's advantage to note that whatever the

skills are that it measures, they are related to some real-world job behaviors. The present measure has been correlated with other tests such as the *Differential Aptitude Tests*, *California Test of Mental Maturity* and *The Wonderlic Personnel Test*. Correlations with these instruments were not particularly high (in the .58-.65 range) indicating not too great an overlap with the skill levels required by the standardized instruments. This fact is advantageous under the criteria of the present review.

c. Construct validity. No information is given in the manual.

2. Reliability.

a. Comparability. The equivalence of the two forms of the test was examined with three separate samples, and reported separately for each component of the test. Although the manual reports that the two forms are similar in content and comparable in difficulty, there is no way of determining whether the number and kind of items are equivalent. Only one form was included in the package. Standard deviations and means are reported statistically equal for the two forms in all three samples, with one exception. This exception disappeared when the three samples were combined. Thus, for

practical purposes, the two forms may be considered equivalent. No correlations, however, were reported in this section.

b. Stability. Test-retest coefficients with form "A" revealed coefficients above .90, with a retest time span of two to three months. Using Form B only, stability was a bit lower (high .70's) but still quite good. Using Form "A" first, then "B", reliability still remained at .86 for the combined verbal plus arithmetic score. The sample size, however was only 39, rendering this coefficient open to possible fluctuation from sample to sample. Overall, stability is good, and quite appropriate under the standards of the present review.

c. Internal consistency. Internal consistency was measured with a sample from a Southern city school system, and it was reported by Race ("White vs. Negro"), and by grade (Grades 6, 8, 10 and 12 reported separately). Kuder-Richardson Formula 20 yielded coefficients above .84 for all grades for both White and Black samples. These coefficients were higher for Form B (above .95) although no breakdown by class and Race were given. For purposes of the present review, these

coefficients were considered appropriate.

d. Standard Error of Measurement. For Form "A" it is reported by grade and by race (Black vs. White), for grades 6, 8, 10, 12. The standard errors of measurement indicate that the test can differentiate groups by grade; i.e., there is almost no overlap of scores between grades. Additionally, it also indicates it can separate the White and Black examinees, the latter having lower means in all grade samples.

3. Test-Item Structure.

Only a sketchy statement about item difficulty is made in the manual. It states that enough "easy" items were included so as to permit most examinees to answer a considerable number of items correctly. This area of test review is fairly critical for the present purposes, since information on item construction and item selection allows a judgment of the representativeness of the entire item pool in reference to actual examinee behavior. Lack of information in this area is considered a serious drawback.

Appropriatness:

1. Instructions. Examinee instructions are recorded on tape. The manual only has instructions

about tape loading, unloading, materials, etc. Thus, a reviewer of this test can not, from the manual, tell whether instructions meet the criteria of the present review.

2. Items. Although it is not clear from item construction procedures if actual behavior was sampled, items do appear relevant and motivating. The propriety of the items is maintained.
3. Format and Procedure. The quality of the paper and layout is good, although facsimilies of actual materials were not always presented. Layout, timing and complexity are acceptable under present standards. Responses are recorded on the test booklet. Scoring is done by hand. This form of scoring is not considered adequate.

Normative Standards

Percentile norms are presented for Verbal, Numerical, and Verbal + Numerical scores, for both forms. The narrative data was obtained for both School groups and Industrial groups. For School Grades 6, 8, 10 and 12, norms are available with Form A. Forms A and B were normed on the various Industrial Groups, presumably adults. There seems to be some geographic representativeness in the Industrial sample, although no data was given on the ethnic, sex, and socioeconomic sampling. The School Groups are brokened down by race ("White vs. Negro"), and only a Nor-

thern and a Southern School was sampled. No item statistics are reported. The norms presented in reference to Form "A" are considered usable for purposes of the study envisioned in the present review. The major drawback is that these norms are anchored solely on the normative sample, and thus are not interpretable with reference to behaviors subsumed within the concept of functional literacy.

Administration

The personnel required for testing and scheduling time are acceptable. However, the conditions of testing require special equipment (tape recorders), and the manual scoring system is not amenable to reliable machine scoring. A package of 100 tests of either form costs $25.00. Scoring keys and manual are $1.70 for a set of two. Instruction cassettes are $8.50 each, two must be purchased if both numerical and verbal tests are to be administered.

Interpretation

Beyond the norms mentioned above, no other aid in interpreting scores is given. Scores are not referenced to particular behaviors or tasks which examinees are capable of performing at difficult score levels. Thus, the scores are anchored solely on the normative samples. To the extent that the normative sample is at variance with the population for which the test is intended, the above norms lose their utility.

Evaluation

This test, as is, cannot be rated appropriate for the Title I Study, primarily because it was intended for, tested, and normed with grades 6, 8, 10, 12 and with various Industrial groups. Additionally, there is no information on item construction, and thus it is impossible to determine whether observations of reading behaviors were used for item generation.

On the positive side, this test includes both computational and reading skills, and it has been widely normed.

Modification of this test to suit the purposes of Title I Study could probably be accomplished, provided that the publisher has item statistics on the remainder of the item pool. The new test, of course, could not be judged by present norms, and thus a clinical pretest would become the absolute minimal requirement for observing how the new test would behave with a sample appropriate to the Title I Study.

National Assessment of Educational Progress in Reading

Released Exercises

Education Commission of the States
300 Lincoln Tower
Denver, Colorado 80203

General Description

The National Assessment of Educational Progress in Reading (NAEP) assessed an assortment of reading skills at four age levels: 9, 13, 17, and 26 to 35 years. The study was concerned with the ability of Americans to read printed materials, and more specifically, ". . . with those reading skills usually taught in schools and with the percentages of Americans who have attained those skills." The total sample included 98,016 people ranging in age from 9 years to young adults. The NAEP exercises were developed around 8 themes: 1) understanding words and word relationships; 2) graphic materials; 3) written directions; 4) reference materials; 5) gleaning significant facts from passages; 6) main ideas and organization; 7) drawing inferences; 8) critical reading. Of the original pool of items, nearly 200 have been released for public use. The results obtained can be examined according to several group characteristics: 1) sex; 2) Black or White race; 3) parental education; 4) geographic region; 5) size and type of community; and 6) age.

Test Background

1. Purpose

The purpose of the NAEP exercises is to assess the percentages of Americans who have attained those reading skills usually taught in schools. These skills were categorized according to the eight themes. The themes are based upon six reading objectives: 1) comprehending what is read; 2) analyzing what is read; 3) using what is read; 4) reasoning logically from what is read; 5) making judgements concerning what is read; and 6) having attitudes about, and an interest in reading.

2. Clarity of Purpose to Examiners

In the NAEP study, the items were administered by professionals who had been trained for the task. It is reasonable to assume that the purpose of the assessment was made clear to them. But since a test package is not available, the clarity of instructions cannot be determined.

3. Compatibility of Purpose and Test Construction

The test was constructed on the basis of five reading objectives formulated by a committee of lay and professional advisors. Each exercise was developed within the framework of these objectives. The eight themes were then used to classify the exercises. Thus,

the purpose and method of test construction are highly compatible.

4. Compatibility of Purpose with Item Sampling

The items were not sampled from actual extra-curricular literacy experiences in the lives of the population of persons who would take the exercises.

Psychometric Quality

1. Validity

a. Content Validity

The exercises were not designed as a test of functional literacy in the sense adopted for the Title I Study of Sustaining Effects. Instead, the purpose of the NAEP exercises is explicitly linked to the academic setting, having essentially the same purpose as an achievement test; i.e., to measure reading skills usually taught by schools.

The stimulus materials do not represent those commonly encountered in real-life reading and computational tasks by children in grades 4 to 8. Computational tasks are not included. Most of the stimulus materials would be more common for older children and adults. Moreover, the relevance of the materials to economically or educationally disadvantaged children is unknown.

The real-life reading and computational behavior domains of children in grades 4 to 8 were not directly addressed by the NAEP exercises. The exercises were developed upon a structure of reading objectives. Although these reading objectives probably overlap to some degree with the real-life behavior domains of the sample of children to be tested in the Title I Study of Sustaining Effects, they also include aspects of reading which are probably not critical to a functional level of literacy. In addition, these objectives were formulated by logical means by a committee instead of being drawn directly from the actual reading domain of children. The individual exercises were developed by a test construction contractor and reviewed for acceptance by NAEP consultants.

Both language and graphic representations appear to be uncommon, not entirely relevant, and too difficult for children in grades 4 to 8. This would be particularly true for a student sample containing a substantial number of economically and educationally disadvantaged children.

Many of the NAEP exercises would reasonably be judged to possess properties representing various socioeconomic functions or benefits. Others do not possess these properties. Socioeconomic importance was used as one criterion for acceptance during the NAEP review of potential items.

As desired, the materials are not referenced to specific program objectives. In terms of criterial objectives, the definition of reading skills used by NAEP does not coincide with the definition of functional literacy adopted for the Title I Study of Sustaining Effects. Thus, no clear link between NAEP exercises and the operative definition of functional literacy is possible.

b. Empirical Validity

Results of NAEP exercises have not been related to other measures taken at the same time. Neither have they been related to other measures taken at later times. No studies were reported in which performance on the NAEP exercises was related to other psychological, educational, or socioeconomic independent variables.

c. Construct Validity

The NAEP exercises were not founded upon psychological, educational, linguistic or educational theory, or related to educational practices.

2. Reliability

a. Comparability

Although a series of separate reading exercise packages were constructed, these were not

c. regarded as alternate forms of the instrument, and statistics of comparability were not developed.

b. Stability

No information on test-retest correlations were available.

c. Internal Consistency

No indication of internal consistency was provided.

d. Standard Error of Measurement

Standard errors of percentage of each response for each item is shown for the total national sample as well as for each demographic grouping.

3. Test - Item Structure

a. Item Construction

The definition of functional literacy employed in the Title I Study of Sustaining Effects does not correspond with the reading behaviors addressed by the NAEP. Therefore, the NAEP item pool cannot be said to be relevant and representative of the operative definition of functional literacy. The NAEP items were constructed by a test development contractor and item sampling performed via the expert judgement of a review panel.

b. Item Selection

Selection was based on the judgement of a committee rather than on observation of actual behavior.

c. Item Difficulty

Evidence of the difficulty of individual released items relative to the total set of items in each form was not available. Item difficulty for various demographic strata were provided for each item.

Appropriateness

1. Instructions

a. Clarity

The vocabulary and syntax of the exercise instructions are sufficiently simple and direct in most cases to be suitable to children in grades 4 to 9. In some cases, however, the instructions are somewhat confusing.

b. Purpose

The released exercises are not accompanied by a statement to examinees on the purpose and intended use of the exercises.

c. Comprehensiveness

The exercise instructions are sufficiently comprehensive in describing the task requirements to the examinee. Although no information on the relation of guessing and scoring is provided, a response category labeled, "I don't know" was provided for many exercises.

d. Sample Items

Because the NAEP released exercises are not formatted into a test package, a sample item was not offered.

e. Mode

Separate instructions accompany the stimulus materials for each exercise. Exercise instructions would be amenable to presentation in an oral mode.

2. Items

a. Motivation

Most items do not appear sufficiently relevant or interesting to inspire the intrinsic motivation of children in the 4 to 8 grade range.

b. Propriety

No invasion of privacy, sexist, racist, or otherwise offensive content was identified in the NAEP exercises.

3. Format and Procedure

a. Physical Quality

The NAEP exercises are not presented in a test package. Therefore, many factors pertaining to physical quality would be the responsibility of the secondary user. The quality of graphic representations, a factor inherent in the exercises, is good.

b. Layout

The arrangement of items permits ready recognition of separate items. However, the relation of the stimulus material to the exercise question is sometimes inadequate for clarity.

c. Timing

Because the NAEP exercises are not formatted into a test package, the secondary user of these exercises would determine the test time by selecting a particular number of exercises for a given administration need. In their primary use, sets of exercises were used which required a 35 minute test period.

d. Response Mode

As presently formatted, the NAEP exercises are not amenable to machine scoring.

e. Complexity

Each exercise requires a single, simple, and direct response. Many of the stimulus materials have been used as the basis for multiple items.

Normative Standards

1. Data Available

The NAEP Study obtained data on each exercise for a variety of demographic variables. The item difficulties are therefore available for children aged 9, 13, 17, and for adults, as well as for race, geographic region, etc.

2. Normative Sample

The available data on NAEP exercises was obtained for children aged 9, 13, and 17 years, and adults aged 26 to 35 years. The primary strata of the sample (geographic region and community type and size) did not systematically address the age and economic variables that are central to the Title I study.

3. Representative

The NAEP sample was not systematically stratified. The results obtained from a survey of 98,016 people were examined on the basis of various group characteristics including: sex, Black or White race, parental education, geographic region, size and type of community, and age.

4. Reporting

The data for the NAEP released exercises are reported separately for a variety of group characteristics.

5. Item Statistics

Item statistics are reported for both the whole sample and for the various group characteristics.

Administration

1. Personnel

The NAEP exercises are amenable to administration by nonexpert personnel.

2. Scheduling

A set of NAEP exercises could be selected to produce a 30 minute testing period.

3. Setting

The NAEP exercises are amenable to group administration in a usual classroom setting.

4. Scoring

Some of the NAEP items are machine scoreable, while others require hand scoring. The reliability of the machine-scoring procedures is unknown.

5. Materials

The NAEP exercises are entirely of the paper-and-pencil variety.

6. Cost

The NAEP released exercises belong to the public domain.

Interpretation

1. Manuals

The NAEP released exercises are not formatted into a test package. Consequently, there is no test manual.

2. Meaning

The meaning of a score on an NAEP exercise is derived purely from the examinee's ability to successfully cope with the content and demands of the exercise. Performance on NAEP items cannot be related to a hierarchy of performance levels.

3. Scales

The primary exercise scores are not directly understandable.

4. Implications

The implications of performance on NAEP exercises is limited to the ability to perform any particular exercise. The implication is related to the NAEP reading objectives to the extent that the exercises are valid derivatives of these objectives. The relationship of NAEP performance to educational policy has not been developed.

Evaluation

The advantages of the NAEP exercises are that they have been used with children of the age group to be involved in the Title I Study of Sustaining Effects, and item statistics are available for several examinee variables that are important for the Title I purposes (e.g., age, race, parent education, geographic region, and community size). Further, they could be formatted into a test package of appropriate length and to meet other criteria of administration.

In spite of the positive qualities, the NAEP exercises have several undesirable features. The most glaring rises from the lack of a computational subtest. In addition, the stimulus materials are not judged to be intrinsically motivating and post pilot work with them has revealed them

to be sufficiently difficult for children in the 4 to 8 age range as to generate some degree of examinee resistance. The item development process has the conceptual difficulty of having been created by experts, rather than having flowed from observations of real-life experiences of the population of potential examinees.

The possible utility of the NAEP exercises to the Title I study is two-fold. First, an examination of the item statistics of certain items judged to have qualities of intrinsic motivation and suitability to the Title I assessment purposes may result in the selection of particular items. Secondly, the NAEP items may have heuristic value to the development or modification of a functional literacy assessment tool. As they stand, however, the collection of items for 9 and 13 year olds do not in the reviewers opinion, represent an acceptable means of assessment for the Title I study. At a minimum, this collection would have to be modified and supplemented extensively.

New York State Basic Competency Test in Reading - A Derivation of the Adult Functional Reading Study by Educational Testing Service

The University of the State of New York
The State Education Department
Albany, New York 12234

General Description

The New York Basic Competency (NYBC) Test in Reading is available in an experimental form only. It consists of 28 different samples of reading materials, each accompanied by one, two, or three questions which are designed to measure the comprehension of students in grades 9 to 12. The total score of 40 is generated by 40 multiple choice questions. On the basis of an arbitrary cutoff score of 65%, a score of 26 is recommended as the minimum passing score. The test requires approximately 45 minutes, but this time period has been implemented in a flexible manner in order that examinees may be allowed more time to complete the test if they need it. The test is group administered by nonexpert personnel.

Test Background

1. Purpose

"The NYBC Tests are designed to provide a measure of the extent to which pupils, beginning in grade 9, have achieved a minimum level of mastery of the basic competencies that will be required of them as adults. The goal of the school is to assure that every pupil will have reached such basic competency levels before

leaving high school, whether as a graduate or a drop-out."

2. Clarity of Purpose to Examiners

No concrete explanation was given in the manual of "basic competencies," nor of the basis for decisions on what is required of adults. A review of the Adult Functional Reading Study performed by Educational Testing Service (ETS) would be necessary (Murphy, 1973) before these aspects of the test's purpose would be understood.

3. Compatibility of Purpose and Test Construction

The New York Basic Competency Test in Reading is a direct derivation of the instruments developed by ETS in the Adult Functional Reading Study (Murphy, 1973 - 1975). Thus, the two primary construction efforts made by ETS in that study also apply to the NYBC Test in Reading; i.e., 1) the nation-wide survey of actual adult reading experiences, and 2) the large-scale field test of instruments constructed on the basis of the survey results. The purposes underlying these two efforts were to determine the nature of the actual reading experiences of American adults, and to develop a means of measuring an adult's ability to cope with these real-world literacy experiences.

The purpose of the NYBC Test in Reading is very similar to the measurement objective of the Adult Functional Literacy study; namely, to measure the extent to which persons have achieved a minimum level of mastery of the reading competency expected of them as adults. The major difference is that the NYBC Test's purpose is targeted for a particular segment of people, secondary students in grade 9 onward. If the ETS tasks are accepted as valid instances of the reading competencies that will be expected of these persons, then the purpose of the New York State Basic Competency Test in Reading is highly compatible with the manner in which it was constructed.

4. Compatibility of Purpose with Item Sampling

The universe of items, from which the test items were sampled, were obtained from a stratified sample of American households. Approximately 100 primary sampling units were generated. These set geographic region and size of community as the first and second order strata from which a representative sample of U. S. households were sampled. One person age 16 and over was selected for interviewing from each household by a predetermined selection table. No special provision was made

Reading/Everyday

for sampling items in a manner which would accommodate differences between ethnic, racial, or varying income groups.

Psychometric Quality

1. Validity

a. Content Validity - The NYBC Test in Reading is defined as a measure of the minimum level of mastery of basic competencies required of adults. This definition may reasonably be interpreted as a measure of functional literacy.

The stimulus materials, by virtue of the method by which they were generated in the ETS study, represent those commonly encountered in real-life reading tasks of persons in grades 9 upward. However, the relevance of the materials to the experiences of economically or educationally disadvantaged persons is unknown. It is known that they were not sampled from children in grades 4 to 8, and therefore do not represent the materials encountered in real life by children of the type involved in the Title I Study of Sustaining Effects. The behavior domain is not matched to tasks and skills required of children between the ages of 9-14.

In selecting items for the test's target group of 9th to 12th grade students, skill and difficulty level were considered by the staff members of the Bureau of Reading of the New York State Department of Education. No explicit criteria for reviewing items on this basis were reported. Apparently, expert judgement was used. However, the second field testing was limited to a sample of students in which 40% fell below the Statewide Reference Point on the PEP test.*

Although the symbolic domain is fairly appropriate with respect to pictorial representations and degree of technical language, the general vocabulary level appears to be too high for children in grades 4 to 8. The content is the most inappropriate aspect of the items. Benefit of task performance was specified as one of the criteria used to select items for the test. This procedure is not described. In the ETS study, a socio-benefit rating was obtained on all items from an advisory panel.

---

*The Statewide Reference Point is the cutoff point between the third and fourth Stanine, with 23% below and 77% above.

It is unknown whether or not the ETS benefit ratings were utilized in the selection of items for the NYBC Test in Reading.

In the judgement of Pacific Consultants, the 40 items in Form L are of relatively high socioeconomic benefit. They do suffer however, from the omission of some areas which could be of greater value to the grade 4 to 8 children of the Title I Study of Sustaining Effects. They are also weakened by the somewhat contrived nature of some items.

As desired, the materials are not referenced to specific program objectives. In terms of criterial objectives, the tasks of the NYBC Test in Reading benefit from the rationale underlying the development of instruments in the Adult Functional Reading Study. Consequently, the definition of functional literacy is based on a large scale sample of the actual reading tasks of adults, and the test items are versions of these tasks. Thus, the link between definition and items is clear. The one reservation is that the ETS items were based upon a survey of persons aged 16 and older, while the NYBC Test is used with persons approximately 14 to 17 years old. These two age groups are in contrast

to the age range of 9 to 14 years to be addressed in the Title I Study of Sustaining Effects. Thus, to the degree to which it is reasonable to assume that the difference in age group modifies the definition of functional literacy, the link between definition and the test items becomes weakened.

b. Empirical Validity - Results are reported on the relationship of this test with the "Ninth grade PEP tests in reading and mathematics." No description or explanation of the PEP tests is provided.

Discussion of results indicated that ". . . the Basic Competency Tests and the PEP tests have a considerable overlap of function in both areas, more noticeably in mathematics. Still, the correlations are well below the level required to consider the Basic Competency Tests and the PEP tests 'parallel'."

Approximately 5% of the students who had PEP scores above the Statewide Reference Point failed the NYBC Test in Reading. Slightly more than half of the students obtaining PEP test scores below the Statewide Reference Point were able to pass the NYBC Tests. No information

was available on the other aspects of empirical validity.

c. Construct Validity - The test is not based directly upon the theoretical constructs of psychology, education, and psycholinguistics; nor upon educational practice and policy.

2. Reliability

a. Comparability - In one report on the test its relation to the PEP test was reported. This report also eluded to parallel forms. Only Form L was made available to Pacific Consultants. No information was available on the construction of parallel forms.

b. Stability - No information on test-retest correlations were available.

c. Internal Consistency - The math test is not available to Pacific Consultants. Neither are data relevant to the relationship between the math and reading scales of the test available.

d. Standard Error of Measurement - not available.

3. Test-Item Structure

a. Item Construction - This process was not clearly described. Aside from the knowledge that the items are derived from the original pool established in the ETS Adult Functional Reading Study, nothing is known about how the items were selected

for the various forms of the NYBC Test in Reading.

b. Item Selection - No information is available.

c. Item Difficulty - Statistics calculated on item difficulty showed that the difficulty level for the items ranged from .35 to .97, and that an average difficulty of .80 was obtained for the eight preliminary forms of the test. Table 1 below shows the number and percent of items for each difficulty level.

Table 1. Number and Percent of Items at Successive Levels of Difficulty.

Level of Difficulty	Number of Items	Percent of Total Items
.90+	48	41
.80 - .89	26	22
.70 - .79	14	12
.60 - .69	12	10
.50 - .59	8	7
.40 - .49	5	4
.30 - .39	3	3

Appropriateness

1. Instructions

   a. Clarity - The vocabulary used was somewhat too sophisticated for children in grades 4 to 8. The requirement that the student read the item instructions contradicts some of the fundamental assumptions about the need to measure minimal competency in reading.

   b. Purpose - The purpose of the test was explained well if the student can read the directions on page 2 of the test booklet.

   c. Comprehensiveness - The comprehensiveness of instructions is satisfactory if the student can read them.

   d. Sample Items - The sample item accurately illustrates the task requirements and the level of task difficulty. These benefits may not apply however, if the narrative surrounding the sample item cannot be read.

   e. Mode - The instructions must be read by the student. This is a major drawback as the reading skills that are needed exceed those necessary to respond to many test items.

2. Items

   a. Motivation - Because the content of the items is relevant only to adults, and not to children, poor intrinsic motivation should be expected for 4th to 8th graders. The contrived nature of some items may also preclude the high intrinsic motivation of older students as well.

   b. Propriety - No instance of invasion of privacy was found in the items. Deliberate attempts to utilize non-sexist, non-racist language and content were apparent.

3. Format and Procedures

   a. Physical Quality - The paper, print, and pictorial representations were of reasonably good quality.

   b. Layout - The test layout clearly separated the items from one another.

   c. Timing - Test length was developed with timing in mind (approximately 1 class period of 45 minutes), but open-ended timing was recommended. These proposed lengths are both in excess of the 20 to 30 minute length desired for the Title I Study of Sustaining Effects.

   d. Response Mode - Five types of answer sheets are available for each test, four of which are intended solely for machine scoring, and

one for either machine or hand scoring.

e. Complexity - Several items are based on some of the test materials. However, each item requires only one simple and direct response.

Normative Standards

No normative data on children in grades 4 to 8 are available. Neither the sampling procedures for item generation, nor those for field testing attempted to obtain a sample representative of racial, ethnic, or socioeconomic strata. Geographic and community size were the primary strata for the item generation task. Field testing was performed in New York State communities of varying size.

Administration

The NYBC Test in Reading can be administered by non-expert personnel. Test timing could easily be modified for a 30 minute period. The test is suitable for administration in a usual intact classroom setting. Scoring is by machine. The reliability of the machine scoring process required is unknown.

The test materials are entirely of the paper-and-pencil variety. The test cannot be purchased from ETS. The New York State Department of Education considers it an experimental instrument which is not yet ready for dissemination.

Interpretation

1. Manuals - The test manual is very brief and does not conform to APA standards for test manuals.

2. Meaning - The meaning of a score on the NYBC Test in Reading is not clear.
3. Scales - The primary test scores are not directly understandable.
4. Implications - The implications of the test score is that the examinee can or cannot perform certain minimal real-world tasks. The relationship of this implication to educational policy has not been fully developed.

Evaluation

The lack of relevance of the items of the New York State Basic Competency Test to children having the age and socio-economic attributes of those in the Title I study, presents a serious drawback for the test's consideration. Further, the test does not contain a computational section. The issue of item suitability is so problematic that it renders the modification of this instrument impractical.

Reading/Everyday Activities in Life (R/EAL)

Marilyn Lichtman, Ed.D.
Virginia Polytechnic Institute

Cal Press Inc.
76 Madison Avenue
New York, New York 10016

Description

R/EAL is a test of reading, divided into nine reading "selections", each of which is claimed to represent a general category of reading, "...often encountered by individuals of high school age or above." The manual also states that "...all indications are that it should be useful with anyone age ten or older."

The nine reading selections, with five questions per selection, are as follow:

1. A set of road signs.
2. A T.V. schedule.
3. A set of directions for preparation of cheese pizza.
4. A reading selection on the topic of narcotic drugs.
5. A food market ad.
6. An apartment lease.
7. A road map.
8. A Want Ad.
9. A job application.

The test is administered by means of individually operated cassette players and earphones, which allows the

test to be self-administered, self-directed and self-paced. Group administration is also possible by having audio equipment for each student. Recently a script of instructions has been made available which would eliminate the necessity of audio equipment.

Test Background

1. Purpose

The manual states that, "the R/EAL should be used to assess whether or not an individual is functionally literate."

It claims a suitability for minorities, Blacks, Puerto Ricans, Mexicans, and others, who have been singled out by the bias of the traditional standardized tests.

As an evaluation tool, R/EAL claims to allow a determination of progress made by students. It can also be used to measure to what extent students in a given program have basic literacy skills. It cannot be determined from the manual whether test construction follows it's intended purpose.

Psychometric Quality

1. Validity

The R/EAL alleges to measure, on an individual basis, responses to questions that are easily identifiable with the examiners' every day life.

The correlation between the R/EAL and the Stanford Achievement Test is .74. Content validity is based upon the generation of items from a task analysis used in the definition of test objectives.

2. Reliability

Using a minority sample of persons with an average of 5th grade reading achievement scores, an internal consistency estimate of .93 was made with the Kuder - Richardson Formula 20.

3. Test-Item Structure

The items in the R/EAL were identified only as reading tasks that one could reasonably expect to encounter in his/her every day encounters. A review of the items reveals that some of them would be inappropriately difficult for the low end of the intended sample (e.g., items based on a facsimili of an apartment lease).

Appropriateness

1. Instructions

Recorded instructions were sufficiently clear and an explanation of the purpose was provided in the manual. The R/EAL provided two sample questions in the test booklet.

2. Items

The recorded administration procedures which permitted self-pacing and self-administration

may have a positive effect on examinee motivation. The use of facsimilies of real-life literacy experiences may have also enhanced the motivation of adult examinees. The irrelevance of these materials to the lives of children in grades 4 to 8 makes this a questionable advantage to the present study. The items were not judged by the reviewers to have offensive content.

3. Format and Procedures

The test booklet was of generally good physical quality, and the illustrations, and graphics were realistic. The printed layout of the R/EAL was adequate. However, since the sequencing was heavily dependent upon the tape recorded instructions, mechanical problems could seriously disrupt the necessary link between instructions and stimulus materials.

The R/EAL has made provisions in its design for the differences in the speed with which a student can finish the test. It also is taped in such a way that individuals and groups can start and finish at different times, without having negative affects on the examinee.

There is no designated amount of time, given to the students to finish the test.

The R/EAL's test booklet is arranged in such a way, that students can write directly in the test booklet, but it is also designed to be hand scored, using a key provided in the manual.

Normative Standards

Normative data is not available. The manual for the R/EAL suggested that the test was highly suitable for minorities, Blacks, Mexicans, Rural groups, and all of those shown bias by the standardized test. However, NO EVIDENCE is presented to substantiate this claim. Item statistics were not presented.

Administration

The R/EAL may be administered in a classroom setting by reading a prepared script of instructions. It can be administered by non-expert personnel. The test requires approximately 20 to 30 minutes. Hand scoring is required.

Interpretation

1. Manuals

The manual for the R/EAL was fairly elaborate and clear in the areas that it covered, but obviously missing was norming, or any evidence for purposes of test interpretation.

2. Meaning

The only clue to score interpretation is the statement that 80% is "passing". There is no information as to how this percentage was deduced.

3&4. Scales and Implications

Beyond the "passing" score level, no information is available to make a judgement on these criteria.

Evaluation

The main strength of this test is the real-life characteristics of its item formats. Beyond this, little can be said to recommend it for use in the Title I Study. There is no evidence as to whether the items can be used with children in grades 4 to 8. No computational subtest is available. The use of cassettes or tapes makes it unusable for present purposes, although the newly prepared instruction script could alleviate this drawback. A passing grade of 80% is the only normative data presented, and is not related to other kinds of real-world performance.

IMPLICATIONS AND RECOMMENDATIONS

As a result of the literature review, and the test evaluation activities of the functional literacy project, a series of conclusions and recommendations have become apparent.

Available Tests Judged by the Review Criteria

None of the tests reviewed satisfactorily matched all criteria. The inadequacies of the tests for purposes of the Title I study are basic. No test could be found that is appropriate to the 4th to 8th grade age group. The tests were constructed either for adults, young adults, or pupils from grade 6 to grade 12. All six of the tests contain varying amounts of materials that are commonly encountered in real life reading, with test items constructed from these materials. However, none of the materials were obtained by actual observation of the behavior of children. Instead, an "expert" judgement was made as to the materials that people would have to read in order to function in society. The problem with this approach is that this kind of "expert" judgement can defensibly be made for adults (i.e., most adults have to pass a test to get a driver's license), but it does not apply to young children aged 9-14 years.

Two of the tests, the Adult Performance Level Test (APL), and the Fundamental Achievement Series (FAS), measured both reading and computational skills.

They cannot be used for the Title I study without modification because the former was designed for adults only, and the latter for grades 6 and above. The FAS could probably be modified at less expense, since it already covers the higher end of the intended sample; i.e., grades 6 and 8. Instructions however, are tape recorded, and would therefore require modification. Although there is no evidence that the test items in either the FAS or APL were built around actually-observed, real-life reading behaviors, the APL items are accurate facsimilies of literacy stimuli commonly encountered by adults. Various other important criteria were lacking in these tests, but since the most basic criteria were not met, it is a moot point to discuss additional inadequacies. Detailed descriptions of these criteria are included in the test reviews.

The remaining tests, which measured reading skills only, have various advantages and disadvantages. The National Assessment of Educational Progress in Reading (NAEP) had the most information on each item. Although individual items rather than an administration-ready test package are actually available, if the construction of a test is envisioned or decided upon, then serious consideration should be given to some of the items presented. Item statistics are available by sex, race, geographic region, size and type of community, and age.

The favorable aspect of the Basic Skills Reading Mastery (BSRM) Test is that it has a test form for children 12 years of age. It therefore was constructed for children who match at least a portion of the Title I age range. The lack of realistic facsimilies as item materials depreciates its value to the disadvantaged population of children to be studied. Rather than having been developed from observations of the actual reading experiences of children, the BSRM materials were based upon expert judgement. Modification of this test would require that a set of easier items be added for 9 to 11 year old children and that the representativeness of stimulus materials be improved. Further, test instructions would have to be converted to an oral presentation mode.

The Reading Everyday Activities in Life (R/EAL) has the important feature of presenting its items as actual photographs, or true-to-life drawings of the objects that contain the reading matter. This is an extremely desirable characteristic of item presentation, especially where minority groups are to be tested. Unfortunately, most of the items were constructed for an intended population of high school graduates and older. Additionally, there is no evidence that actual reading behavior was used as the basis for item construction. Modification of the R/EAL for use in the Title I Study would only succeed if certain items were selected which are judged to have relevance to 4th to 8th grade children, and then supple-

mented with other easy and appropriate tasks. The instruction script and other details of administration would require minor modifications in order to make them entirely compatible with the testing purposes of the Title I Study.

Possible Courses of Action

Since no test actually exists which meets the needs of the Title I Study, two courses of action other than test selection, are possible.

1. The most idealistic solution would be to develop a test from the beginning. This would entail sampling the actual reading behaviors of children who match the Title I age and demographic characteristics. A technically sound item-building phase would then be required, with careful pretesting of the final instrument. The advantage of this option is that the final instrument would be a high-quality test which would be suitable for a wide range of future applications. Moreover, it would be most responsive to concerns regarding the testing of disadvantaged children. The disadvantage lies in the time and resources required, since it would be the option requiring the longest time-table. The work could not be completed within the present scope of work, nor by the December 1, 1975 deadline for submitting an instrument package to OMB for clearance.

Should accommodations be made for the time and level of effort required, Pacific Consultants has the capability required to implement this option.

2. The second option open to the Title I Study's assessment of functional literacy is the construction of a test from multiple sources. This procedure would entail the combination of a computational section with either a partially suitable test of reading competency, or with a reading section constructed of items drawn from several instruments. In either case, some set of easy items would have to be added to both the reading and computation portions of the assessment tool. These easy items could either be new creations, or more likely, modifications of items extant in some of the items available from the six tests reviewed.

As an example of this second option, the BSRM, published by the State of Maryland, may be modified so as to include items appropriate for younger children. Then, the computational items of the FAS could be used to construct a computational subtest. Alternatively, items from the R/EAL, or items from the NAEP could be modified or selected for use as reading

tasks, and used in combination with the computational items from the FAS. Another version of this option would consist of constructing a test from all possible sources, sampling and building items by the use of test construction "experts." Any test produced by these methods would then have to be pretested, and modified at least once before it would be ready for a field pre-test. Pacific Consultants has the staff and technical capability to carry out this option within the current time-table and level of effort. Although this approach precludes the advantages of generating literacy tasks from the actual experiences of disadvantaged 4th to 8th grade children, it does maintain the qualities of a criterion-referenced approach to instrument construction. In so doing, it would supply the Title I Study of Sustaining Effects with a suitable instrument to compliment the norm-referenced standardized tests of achievement.

REFERENCES

American Psychological Association. Standards for educational and psychological tests. Washington, D.C.: Prepared by a joint committee of American Psychological Association, American Educational Research Association, National Council on Measurement in Education, Frederic B. Davis, Chair, 1974.

Center for the Study of Evaluation, U.C.L.A. CSE Elementary School Test Evaluations. Los Angeles: Ralph Hoepfner and The Staff of the School Evaluation Project (Guy Strickland, Gretchen Stangel, Patrice Jansen, Marianne Patalino), 1970.

Murphy, Richard T. Adult Functional Reading Study, Final Report. Princeton, New Jersey: Educational Test Service, 1973.

Murphy, Richard T. Adult Functional Reading Study, Supplement. Princeton, New Jersey: Educational Test Service, 1975.

Northwest Regional Educational Laboratory. Tests of functional adult literacy: an evaluation of currently available instruments. Portland, Oregon: Nafziger, Dean H., Thompson, R. Brent, Hiscox, Michael D., and Owen, Thomas R., 1975.

Pacific Training and Technical Assistance Corporation. Evaluation of the Community Based Right to Read Program. Berkeley, California, 1974.

"END OF DOCUMENT"